DOCUMENT RESUME

ED 465 599 SE 066 297

AUTHOR Lawrenz, Frances
TITLE Evaluation of Teacher Leader Professional Development Programs.
PUB DATE 2001-00-00
NOTE 22p.; In: Developing Teacher Leaders: Professional Development in Science and Mathematics; see ED 451 031.
AVAILABLE FROM ERIC Clearinghouse for Science, Mathematics, and Environmental Education, 1929 Kenny Road, Suite 200, Columbus, OH 43210-1080. Tel: 800-276-0462 (Toll Free); Fax: 614-292-0263; Web site: http://www.ericse.org.
PUB TYPE Reports - Descriptive (141)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; *Faculty Development; *Program Evaluation; Teacher Improvement; *Teacher Leadership

ABSTRACT

This chapter outlines the considerations necessary in planning and conducting evaluations of teacher leadership programs. It contains a discussion of different definitions of evaluation, their underlying philosophies, and their importance for teacher leadership programs. Specific ideas for conducting teacher leadership program evaluations are presented. These included: evaluation of the delivery of professional development programs, effects on teacher leaders, effects on classrooms and students, effects within schools, and effects within districts and states. Suggestions include both qualitative and quantitative approaches, and examples are provided from existing teacher leadership programs. The Horizon Research Inc. forms for classroom and professional development observation are described in some detail. (Contains 27 references.) (Author/MVL)

Reproductions supplied by EDRS are the best that can be made from the original document.

ED 465 590

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

14 Evaluation of Teacher Leader Professional Development Programs

Frances Lawrenz
University of Minnesota

This chapter outlines the considerations necessary in planning and conducting evaluations of teacher leadership programs. It contains a discussion of different definitions of evaluation, their underlying philosophies and their importance for teacher leadership programs. Specific ideas for conducting teacher leadership program evaluations are presented. These include: evaluation of the delivery of professional development programs, effects on teacher leaders, effects on classrooms and students, effects within schools, and effects within districts and states. Suggestions include both qualitative and quantitative approaches and examples are provided from existing teacher leadership programs. The Horizon Research Inc. forms for classroom and professional development observation are described in some detail.

The purpose of this chapter is to outline the considerations necessary in planning and conducting evaluations of teacher leadership programs. The chapter begins with a look at the various definitions of evaluation and their relationships to different evaluation philosophies. Next the importance of evaluation for teacher leadership programs is discussed. After the introductory sections, specific ideas for conducting evaluations to determine effects of teacher leader programs are presented. These evaluation methods are grouped under different types of effects and include evaluation of the delivery of professional development programs, effects on teacher leaders, effects on classrooms and students, effects within schools, and effects within districts and states.

What is Evaluation?

The Joint Committee on Standards for Educational Evaluation (1994) defines evaluation as the systematic investigation of the worth or merit of an object. Objects include educational and training programs, projects, and materials. Michael Scriven in his *Evaluation*

SE 066 297

BEST COPY AVAILABLE

Thesaurus (1991) agrees with this definition and goes on to say that the process normally involves: "some identification of relevant standards of merit, worth, or value; some investigation of the performance of evaluands on these standards; and some integration or synthesis of the results to achieve an overall evaluation" (p. 139). One of the first definitions of educational evaluation was provided by Daniel Stufflebeam and the Phi Delta Kappan National Study Committee on Evaluation in *Educational Evaluation and Decision-Making* (1971). In this book the authors say "the purpose of evaluation is not to prove but to improve" (p. v). They define evaluation as the systematic process of delineating, obtaining and providing useful information for judging decision alternatives. This definition is particularly useful in that it highlights that evaluation includes determining what type of information should be gathered, how to gather the determined information and how to present the information in usable formats.

Michael Quinn Patton in his book *Utilization-Focused Evaluation* (1997a) reiterates and expands on the notion of usefulness by making it clear that the receivers of the evaluation information need to be substantively involved in the evaluation process so that the resulting information will be used effectively. A recent addition to the definitions is David Fetterman's (1996) empowerment evaluation. There has been considerable debate about this approach. Fetterman (1997) describes empowerment evaluation as a shift from the previously exclusive focus on merit and worth alone to a commitment to self-determination and capacity building. In other words, empowerment evaluation is evaluation conducted by participants with the goal of continual improvement and self-actualization. Patton (1997b) places empowerment evaluation into a larger context of emancipatory research and goes on to say that teaching evaluation logic and skills is a way of building capacity for ongoing self-assessment. Emancipatory research is the process of using research to improve the researcher and provide the capacity for even more sophisticated self-knowledge and self-determination.

How are the Definitions of Evaluation Related to Evaluation Philosophies?

The different definitions and models for evaluation are based in different philosophies. House (1983) has categorized these differing philosophies along two continua: the objectivist-subjectivist epistemologies and the utilitarian-pluralist values. Objectivism

requires evidence that is reproducible and verifiable. It is derived largely from empiricism and related to logical positivism. Subjectivism is based in experience and related to phenomenologist epistemology. The objectivists rely on reproducible facts while the subjectivists depend upon accumulated experience. In the second continuum, utilitarians assess overall impact while pluralists assess the impact on each individual. In other words, the greatest good for utilitarians is that which will benefit the most people while pluralism requires attention to each individual's benefit. Often utilitarianism and objectivism operate together and pluralism and subjectivism operate together although other combinations are possible. These combinations lead to a wide variety of evaluation approaches and methods. Given the definitions above, Scriven and Patton would be in the middle of the road; Fetterman would be nearer the subjectivist and pluralist poles; and Stufflebeam would be nearer the objectivist and utilitarian poles.

Another elaboration on evaluation is necessary. Although there are many similarities, evaluation and research are not the same and their uniqueness should be kept in mind. Worthen, Sanders and Fitzpatrick (1997) describe the distinction quite well. They point out that evaluation and research differ in the motivation of the inquirer, the objective of the inquiry, the outcome of the inquiry, the role played by explanation, and in generalizability. Evaluators are almost always asked to conduct their evaluations and therefore, are constrained by the situation. However, although researchers may apply for grants to conduct their research, they are generally the ones that make the decisions about why and how to conduct it. The objectives and outcomes in the two types of inquiry are also slightly different. Research is generally conducted to determine generalizable laws governing behavior or to form conclusions. Evaluation, on the other hand, is more likely to be designed to provide descriptions and inform decision making. Finally, evaluation is purposefully tied to a specific object in time and space while research is designed to span these dimensions. These distinctions are important because they affect the type and appropriateness of evaluation designs. Because of their tie to specific situations, evaluations are both less and more constrained than research. They are less constrained because they do not have to be universally generalizable, but they are more constrained because they have to fit into a specific context.

Each of the different approaches to evaluation has its own strengths and limitations, so careful selection of approaches is critical

(National Science Foundation, 1997). Approaches need to be tied to the uses that will be made of the evaluation information by the various audiences. In teacher leadership programs there are many different audiences for evaluations. A potential list includes: the teacher leaders themselves, the people running the professional development programs, the classroom students of the teacher leaders, the colleagues of the teacher leaders, the schools and districts in which the leaders work, the state and national organizations interested in professional development and student learning, and agencies funding the programs. Each of these audiences has its own special information needs and may respond differently to different types of evaluative data. An evaluator should identify these needs, preferences and potential responses as the evaluation is being planned so that the resulting data will be used most effectively. Thoughtful analysis, sensitivity, common sense and creativity are all needed to make sure that the actual evaluation provides information that is useful and credible (Stevens, Lawrenz & Sharp, 1993).

Why is Evaluation Important for Teacher Leadership Programs?

Evaluation can meet several needs in the professional development of teacher leaders. First, evaluation can provide information that helps to justify the program. This type of information is of most interest to program planners and to program funders. In Stufflebeam's (1971) CIPP (Context, Input, Process and Product) model this type of evaluation would be in the context and input realms. The evaluator would be determining what the needs are for a program of this type among the constituents, which would guide the choice of objectives and assignment of priorities (context evaluation). Additionally, given the constraints and opportunities in the situation, the evaluator would be determining what mechanisms would be most feasible (input evaluation). Although both of these types of evaluation can be used in summative and formative fashions, these generally are formative in nature. They help a program decide what to do and how to do it. Another way to think about these types of evaluation is to envision them as checking on the logical contingencies of the program plan (Stake, 1968). In other words, is it logical to expect that the procedures the program is proposing will produce the desired outcomes given the potential participants? Typical questions are:

1. What needs are there for this program?
2. Who are the stakeholders?
3. Are the goals appropriate?
4. What is the best way to accomplish the goals?
5. Will these procedures fit within the situation?
6. Given these potential participants, should these procedures be effective?
7. Is it logical to expect these outcomes after treating the participants in these ways?

Another reason for evaluation is accountability. This type of evaluation is generally both summative and formative. A summative evaluation approach helps program planners know if they are accomplishing their goals. This demonstrates if the program is successful or not to various stakeholders. A more formative approach would be determining strengths and weakness in the program or the leaders it produces. This type of information helps the program to improve itself. Accountability information can be both process and outcome oriented. In terms of process, an evaluator can examine how a program operates, how its procedures combine, and how effective they are. In terms of outcome, an evaluator can determine the effects of the process on the teacher leaders. Typical questions for summative and formative evaluation are:

1. Are they doing what they said they were going to do?
2. Are effective management structures in place to support teacher leaders?
3. Are communication channels open and operating between teacher leaders, teachers, and school administration?
4. Are goals understood and shared by all?
5. Are the presenters in the professional development sessions well qualified?
6. Are the sessions well planned?
7. Do the participants believe they have benefited from the sessions?
8. Do the participants expect to change their behavior?

9. Has the behavior of the participants changed?
10. Have other teachers or students benefited from the changed behavior of the participants?
11. Have schools been affected?

Finally, effectively used program evaluation can instill in the teacher leaders belief in the usefulness of evaluation. The outcome of evaluation is enhanced by the inclusion of empowerment evaluation techniques where the participants in the program are intimately involved in the evaluation effort. Participant involvement with the process of evaluation helps to align the goals of the evaluation with the goals of the teacher leaders. It also provides teacher leaders with evaluation skills to use in other settings. In other words, substantial involvement in evaluation of the professional development program is an excellent opportunity for extending the professional development of the teacher leaders. In order to accomplish this, the teacher leaders must be given the power to determine at least part of the evaluation effort. Teacher leaders should specify program goals they value and determine what data needs to be gathered in order for them to decide if the program has been effective in meeting its goals. Involvement in goal formation helps to make teachers more committed advocates for the program and provides more in depth understanding of program goals. Teacher leaders also need to be involved in data gathering efforts so they will better understand the relationships between goals, data, and decision making. Participants should be able to suggest mechanisms to change the program if their data show it to be ineffective. Empowerment evaluation is most often iterative and incremental. The teacher leaders would specify near term and local goals, determine when these goals were met, and then specify new goals. An analogy for this would be embedded classroom assessment planned by students.

What are Some Evaluation Methods to Determine the Effects of Teacher Learning Programs?

The first three sections provided definitions of evaluation related to evaluation philosophies, justifications for conducting evaluations, and potential evaluation questions. The following section provides specific suggestions for conducting an evaluation by describing various settings and data sources that are possible for information gathering. The examples are intended to be suggestive of various

methods and are not an exhaustive set. These suggestions are based in different evaluation approaches and move out from first order effects to more general and wide spread effects (See Figure 1).

Evaluation of the Delivery of Professional Development

Evaluations of the delivery of teacher leader development can be both formative and summative. Formative evaluation is most useful in situations where similar sessions will be offered in the future. Evaluation can then provide valuable suggestions for improving these future sessions. If more sessions are not offered, evaluation can perform a summative function by documenting the quality of the sessions and their outcomes. Four common immediate techniques for either formative or summative evaluation of professional development are observations, participant opinion surveys, pre post testing of changes, and embedded participant participation.

Some powerful tools have been developed recently by Horizon Research, Inc. (HRI) (1999) for their evaluation of the National Science Foundation's Local Systemic Change (LSC) projects. The LSC program was designed to broaden the impact, accelerate the

Type of Effect	Evaluation Method
Delivery of Professional Development	Observations Participant Opinion Surveys Pre Post Testing Embedded Participant Participation
Effects on Teacher Leaders	Pre Post Testing of Changes Phenomenological Studies Discourse Content Analysis
Effects on Classrooms and Students	Ethnographies Assessment Within Classrooms Assessment of Student Outcomes
Effects Within Schools	Case Studies and Ethnographies Pre Post Testing
Effects on Districts or States	Student Outcomes Policy Analysis Network Analysis

Figure 1. Methods of Evaluation Useful in Determining Different Effects of Teacher Leadership Programs

pace, and increase the effectiveness of improvements in science and mathematics education at the K-12 level. The expectation is that teacher enhancement efforts, standards-based curriculum, parents, informal science and mathematics education institutions, local businesses and industries, nearby colleges and universities and local policies will all come together to achieve a common goal. The program is a mix of federal and local funds with the federal funds supporting teacher professional development, generally through the development of teacher leaders.

A unique aspect of the LSC program is its attention to evaluation. All of the LSC projects are required to participate in a nation-wide evaluation effort, termed the core evaluation, as well as individual evaluation efforts specifically related to local project goals. The core evaluation requirements are: to observe 5-8 professional development sessions per year, to administer 300 teacher questionnaires to teachers in the participating school districts, to administer principal questionnaires to all principals in participating districts, to conduct a minimum of 10 classroom observations, to conduct interviews with 10 randomly selected teachers, and to interview the project administrative team. To ensure uniform data collection, all of the requirements are supported by protocols, surveys or observation formats and evaluators are required to attend national sessions on the appropriate use of the instruments. All of the instruments are available through the HRI home page, www.horizon-research.com.

Observations

One common form of immediate evaluation of teacher leader professional development is observation of the sessions by experts. These observations should use protocols to guarantee comprehensiveness and consistency of the findings. The HRI Professional Development Observation Protocol is an excellent tool (Horizon Research, Inc. [HRI], 1999). It has several components including pre and post interviews with the professional development facilitator and a comprehensive observation protocol that is to be filled out after observing the session for a significant amount of time. The procedure is to interview the presenter, watch a significant portion of the session perhaps taking field notes, interview the presenter again, and then at a later time fill out the Professional Development Observation Protocol (HRI, 1999) using the interview results and the field notes.

The Observation Protocol begins by requesting information about the observer and the contextual background of the session such as the numbers and types of people attending and the focus of the session. Additionally, the observer is asked to categorize the activities the participants are engaged in as listening, reading, discussing, or other activities. After this contextual information the observer is asked to provide various types of ratings. The ratings move from more specific to more general with synthesis ratings at several stages. The synthesis ratings are not intended to be numerical averages of the individual ratings but instead are to represent a holistic impression of program quality.

The observer is first asked to rate the design, implementation, content (science or mathematics pedagogy and leadership), and culture of the session. These ratings include 5 point Likert scales of specific topics. The six items under leadership content are: 1) information on principles of effective staff development was sound and appropriately presented and explored, 2) information on strategies for mentoring and coaching peers was sound and appropriately presented and explored, 3) information on how to be a reform advocate at school or district level was sound and appropriately presented and explored, 4) facilitator(s) displayed an understanding of leadership concepts, 5) participants were intellectually engaged with important ideas relevant to the focus of the session, and 6) participants were given adequate and appropriate opportunity to consider how the content of the session applied to their particular leadership roles. Once the specific topics are rated, the observer is asked to provide a synthesis rating for that portion of the session. These ratings range from 1 to 5. A "1" for leadership content would be "leadership content not at all appropriate for preparing participants to be school or district leaders of mathematics or science education." A "5" for leadership content would be "leadership content highly appropriate for preparing participants to be school or district leaders of mathematics or science education." As additional examples, a "1" for design would be "design of the session not at all reflective of best practice for professional development." A "5" for science content would be "science content of session extremely reflective of current standards for science education." The synthesis ratings are followed by a space for open-ended responses from the observer to provide anecdotal, supporting evidence for the ratings.

After rating these categories, observers are asked to provide overall five point ratings of the likely impact of the session on the

BEST COPY AVAILABLE

participants' capacity to provide high quality mathematics or science education (seven items) and leadership capacity (ten items). The leadership capacity items include the impact on the leaders' knowledge and understanding of mathematics and science, classroom practice, effective classrooms, prior knowledge of teachers, adult learners, the reform process strategies for reform, ability to plan professional development, confidence and networking. Following these sections, there is room for anecdotal comments.

At the end of the protocol the observer is asked to provide a single holistic rating for the overall session. A "1" is ineffective professional development. This is described as, "There is little or no evidence of participant thinking or engagement with important ideas of mathematics or science education." Session is unlikely to enhance the capacity of participants to provide high quality mathematics or science education or to be effective leaders of mathematics or science education in the district. A "5" is exemplary professional development which is highly likely to enhance participants capacity.

An example of a LSC grant project involves the Minneapolis Public Schools (Dr. Carol Johnson, Project Investigator). This project uses this form to evaluate their professional development efforts. One professional development session involved the designated lead teachers from various schools. These teachers spent a week defining important educational issues and studying research dealing with these topics. The facilitators were well prepared, supportive and knowledgeable. Teachers discussed their findings in small, similar interest groups and with the larger group. Overall, this session was given a "4", "accomplished, effective professional development". It was not given a "5" because the evaluator felt it did not adequately address how the teachers would use this information to lead others at their schools. The interviews with the facilitators revealed that they believed they were modeling the behavior the lead teachers would use in their schools. The evaluator felt, however, that more explication of the techniques being modeled and more practice with them would be necessary for it to be "highly likely" that the teachers would be able to use them effectively.

Participant Opinion Surveys

Another type of immediate evaluation is participant opinion surveys, which can use written or oral formats. These types of surveys are designed to gather information about the beliefs of the participants.

BEST COPY AVAILABLE

Participants are asked questions about the worth of the sessions to them, whether their expectations were met, and what might be done to improve the sessions. These surveys are most effective when they are short and collected at times when almost everyone's response can be obtained. Mailing in opinion surveys usually results in a low response rate. It is best to collect them as participants are leaving the professional development program. In sessions of several days, it is useful to have opinion surveys collected half way through so that the remainder of the session can be redesigned to better meet the needs of the participants. The surveys are also more effective when they contain a mix of rating items that target the attainment of specific goals along with a very small number of open-ended questions addressing the issues believed to be most controversial, most ambiguous or the most difficult to express as ratings.

The HRI survey is set up in an oral format. However, some of the items on the HRI Teacher Interview Form (HRI, 1999) could be formatted into questionnaire items. Interviews provide more in depth information but because of the large time constraints, data are gathered from only small numbers of participants. The HRI interview questions include: How do you feel about the professional development? What has been most helpful to you? What has been least helpful? How has the professional development affected you and your teaching? What else do you need to continue improving?

Another technique for obtaining opinions is the focus group (Krueger, 1994). In this technique, a trained focus group facilitator leads a carefully selected group of about 8-15 people in discussions of a small set of provocative questions. This interview technique is widely used in market research where groups of people are asked to try out a new product and then talk about it with each other. The advantage over individual interviews is that you can ask questions of several people at one time, which increases the sample size. Focus groups also provide the opportunity for interaction among respondents that is missing in both written surveys and individual interviews. This interaction helps the facilitator gauge the depth and consensus of feeling about the topics being discussed.

Pre Post Testing

A third type of immediate evaluation is pre post testing of changes in various targeted variables. Examples of pre and post testing variables include knowledge of leadership techniques, knowledge of

other content, perceptions of self as a leader, feelings of empowerment or capacity to lead, or reported past and perceived future behavior. This type of evaluation is more summative in nature and outcome oriented than the participant opinion data described previously. It is also more complicated. In order to document that the participants have changed in a significant way, the instruments for measuring the change must be valid and reliable. There are some instruments that exist and can be used, such as, attitudes toward science, locus of control, personality indices, and understanding of science. Often, however, the goals of the program do not fit exactly with the existing instruments. In this case, new instruments may have to be developed with the concomitant pilot testing to establish feasibility, reliability and validity.

An example of this type of evaluation would be the Physics: A Modeling Approach Project (Dr. David Hestenes, Project Investigator, Arizona State University, Tempe, Arizona) which uses the Force Concept Inventory (Hestenes, Wells and Swackhamer, 1992). The instrument is used in a pre post fashion to determine if future teacher leaders changed their understanding about forces and motion during their professional development. Because the teacher leaders are to be teaching others about science concepts, it is important to know their levels of understanding. The measure provides both formative feedback in the sense of need for more professional development on specific areas of force and motion and summative feedback in the sense of how effective the session was in changing teacher leader understanding of these concepts.

Embedded Participant Participation

A final suggestion would be to use embedded participant participation. This method would involve the participants in the specification of goals for the session, mechanisms for achieving the goals and the designation of data that would demonstrate whether or not the goals were met. Because of the novice status of the participants in terms of evaluation, this could result in a less rigorous evaluation but the process would have the advantage of providing professional development simultaneously. In this case, an evaluation specialist can be used to help coach the participants. Care must be taken with the coach, however, so that the role is indeed coaching. Modeling this sort of coaching behavior is also a valuable source of professional development for the teacher leaders, since they may be required to act in this capacity in their own schools.

In the Minneapolis LSC project the lead liaison teachers are being coached in evaluation and at the same time in using evaluation as a mechanism for helping teachers in their designated schools to move forward in implementing the National Research Council's Science Education Standards (National Research Council, 1996). Meetings are structured where the liaison teachers discuss what they want to accomplish and how they and the evaluator would be able to help determine if their goals are being met. After these data are gathered, the group meets again to discuss the results and determine the next steps. In the schools, the liaison teachers form groups of teachers and together they discuss the best ways to move the school forward in meeting the standards and how they will know when they get there. These planning sessions help to clarify the goals and outcomes to the teachers and allow for their input.

Effects on Teacher Leaders

Pre Post Testing of Changes

The next step away from evaluating the professional development session itself is to examine the effect of the session on the teacher leaders. This provides evidence of the outcomes for the session.

The pre post testing described previously is a measure of the immediate effect of the session rather than a more long-term effect. The pre post testing could be also expanded to include a post-post test where the residual effects of the session would be ascertained. The same test can be used in all three situations. This type of testing can also show the moderating or enhancing effects of experience. Without the initial post test an evaluator would not know if pre to post-post changes were due to the session or to other factors. If there is no pre to post change but there is pre to post-post change, it may be difficult to attribute the change to the session(s). A more sophisticated quantitative design might include repeated measures (Howell, 1987) or time series analyses (Norusis, 1994).

Phenomenological Studies

Another way to study the effects on the teacher leaders would be to conduct phenomenological studies of the lived experiences of some of the teacher leaders. These types of studies are not generalizable across individuals but they do provide rich information about how the professional development impacted the life of the teacher leader. These types of studies take a great deal of time and effort but

their advantage is that they do not force the experience into the narrow categories assessed in pre post and post-post testing. This type of evaluation is an excellent opportunity for the participants to be involved in the evaluation process. One approach may involve teachers keeping their own reflective journals and reviewing each other's experiences. This would not only help to consolidate the experiences through grounded theory but also spread information among the participants about what to do and what to avoid.

Discourse Content Analysis

Another way to study effects on teacher leaders is to analyze their conversations. Discourse content analysis (Kintsch, W., & van Dijk, T.A., 1978; Trabasso, T., van den Broek, P., & Suh, S., 1989) can be done during subsequent meetings and is used as a non-intrusive way to learn what issues are important to them and how they have responded. The major limitation is that the teachers may not talk about issues of importance to the evaluation. It is also difficult to use this type of analysis to make definitive statements about effects. Just because the teachers do not talk about something does not mean that they are not thinking about it. Also, something very important may be mentioned only once while irritating or minor things may be discussed at length. The analysis must proceed carefully and make suggestions not conclusions. This technique is particularly effective if the teacher leaders are part of an electronic communication system. The email discussions can be randomly sampled and a discourse content analysis can be conducted. Having a built in "transcription" of the conversation is invaluable for analyses. The Wisconsin Academy Staff Development Initiative (WADSI) (Dr. Julie Stafford, Project Investigator, Chippewa Falls, Wisconsin) has used the technique to monitor its teacher leaders and it has proven quite informative. Often inferences made from the discourse content analyses are verified with more quantitative survey techniques both on line and on paper.

Effects on Classrooms and Students

These types of evaluations would only be conducted if the program were claiming to have effects on classrooms. It is common in teacher leader programs to assume that the leaders will go back to their school or district and lead other teachers in reform efforts. This assumption would lead to the expectation of change in the classrooms of teachers led by the teacher leader, as well as change in the classrooms of the

teacher leaders themselves. On the other hand, effects on classrooms and the students in them may be too far removed to be attributed to the professional development program. Students in particular are significantly affected by contexts other than school and therefore changes in their behavior are difficult to attribute to any program. Although many types of studies are possible, only three broad categories are discussed here: ethnographies, assessment within classrooms and assessment of student outcomes.

Ethnographies

Ethnographies of classrooms provide the richest data about how the professional development sessions have affected the classroom. Ethnographies are in depth descriptions about the participants, activities, context, and culture operating in particular settings. (Fetterman, 1989). Qualitative techniques are particularly useful for identifying unanticipated effects and in exposing the complex ways in which professional development can lead to change. This is an area where the teacher leaders themselves might be responsible for gathering the data. The most difficult part of qualitative studies is making sense out of the large amount of data gathered. In this scenario, the teacher leaders could be gathering data as they lived the experience and a skilled evaluator could help them make meaning perhaps through a series of focus groups.

Assessment Within Classroms

Assessment of teacher behaviors would require some sort of standard or comparison group against with to compare the behavior of the affected teachers. The science and mathematics standards could be used to formulate behavioral outcomes and then teacher growth on the stipulated behaviors could be measured. Comparison groups could be formed from teachers in schools that did not have teacher leaders. Then the behavior of teachers in one setting would be compared to the behavior of teachers in the other. In order for comparison groups to be effective, careful matching must occur. In teacher leader settings in particular, care must be taken to ensure that the schools and teachers are comparable to ensure that no selection bias exists. It is often the case that schools, or teachers, who are "good" to begin with, will be the ones choosing to participate in professional development. The pre post testing can be quite varied. It could include observations of pedagogy by external or internal "experts" or peers, content analysis

of curricular materials and assessment devices, or student perceptions of teacher activity.

If observations are conducted to assess behavior, a careful protocol should be followed. Again, HRI has developed a Classroom Observation Protocol (HRI, 1999) that can be used to assess the effectiveness of science and mathematics classes. The Classroom Observation Protocol is very similar to the Professional Development Protocol (HRI, 1999) in format. It contains pre and post interviews with the teacher, contextual questions, individual topic ratings, synthesis ratings, an overall holistic rating and the opportunity to include anecdotal evidence in support of the ratings. Once again, the observer is expected to interview, observe, interview again, and then fill out the protocol.

The individual topics on the HRI instrument are grouped into design, implementation, mathematics or science content, and classroom culture. Each of these is given a five-point synthesis rating as well. Then the likely impact of instruction on student understanding of mathematics and science is rated, followed by a holistic rating of the overall lesson. Level 1 lessons are categorized as ineffective instruction, meaning there is little or no evidence of student thinking or engagement with important ideas of mathematics or science. Instruction is unlikely to enhance students' understanding of the discipline or to develop their capacity to successfully "do" mathematics or science. Level 5 lessons are categorized as exemplary instruction, meaning instruction is purposeful and all students are highly engaged most or all of the time in meaningful work; the lesson is well-designed and artfully implemented, with flexibility and responsiveness to students' needs and interests; instruction is highly likely to enhance most students' understanding of the discipline and to develop their capacity to successfully "do" mathematics or science.

Classroom effects can also be assessed through determination of the classroom psychosocial environment. The most common way of assessing this is through a written form that students complete about how they feel about their classroom. Classroom learning environments have been shown to be related to positive student outcomes and to be sensitive indicators of differences in classrooms (Fraser, 1994). One recent form that is aligned with the standards is the Constructivist Learning Environment Survey (CLES) (Taylor, Fraser, & White, 1994).

Assessment of Student Outcomes

Pre post and post-post testing of student outcomes can also be used to assess the effects of professional development. This type of assessment also requires a standard or comparison group to assess against. Student cognitive, attitudinal or behavioral outcomes can be assessed. There is a myriad of instruments available to assess student outcomes. Two reasonable and recent sources for achievement items are the National Assessment of Educational Progress (NAEP) and Third International Mathematics and Science Study (TIMSS) released items (www.nces.ed.gov). Using these items will allow a program to tie their students' achievement to national and international achievement levels. A critical issue in student achievement, attitude, and habits of mind and behavior assessment is deciding when is the most appropriate time to conduct the testing. In order to make this decision, one must decide when change should first begin to appear and how long it should be sustained. There is evidence that you need 2-5 years of implementation to get change in student achievement (Newman, 1996). There is also evidence that student and teacher outcomes are diluted through the "train teachers to train other teachers" approach (Lawrenz, 1986). The first cohort experiences the most significant effect and that decreases as you move out. This dilution, however, does not seem to be the case in the more recent teacher leader professional development models where the leader is directly involved in school or district based planning and development (WADSI Program, Dr. Julie Stafford, Project Investigator).

Effects Within Schools

Many recent teacher leader programs assume that the leaders will return to their districts or schools and become change agents that will stimulate and direct changes at the school level and thereby disperse and increase the effects of the original program. Therefore, examination of school effects is critical. Two qualitative and two quantitative techniques are suggested in the following two sections. A good reference on the effects of school reform, which contains quantitative and qualitative results, is Newman's (1996) study of restructured schools.

Case Studies and Ethnographies

The two most promising qualitative techniques are case studies of schools (Yin, 1989) and ethnography of the culture of the school

(Fetterman, 1989). The ethnography would be used to document the cultural changes that occur as a result of the actions of the teacher leader. This is similar to an anthropologist studying the effects on a native culture when a prominent tribal member returns from an encounter with another culture. The other methodology would be a case study of the school or schools and could include schools with and without teacher leaders or schools utilizing various types of professional development. The strength lies in the fact that several voices and perceptions can be included in the case study. Furthermore, the description is holistic and relates to the entire school. Case studies generally require a long-term relationship with the school and include observations, interviews, surveys, and collection of artifacts.

Pre Post Testing

There are two quantitative techniques that involve pre post testing. One technique is pre post and post-post testing of the culture of the school using surveys. Triangulation of the assessments is through the administration of surveys to three different information sources such as principals, teachers, and students (Louis, Marks & Kruse, 1996). The second technique is pre post testing of student outcomes. This would require that the program assumes that it will have some effect on student outcomes and has all the limitations associated with the effects on classrooms and students.

Effects Within Districts or States

These types of effects are a long way removed from the professional development of teacher leaders but they are often claimed as potential outcomes from these types of programs. There are three different methods that are most useful in determining these effects: student outcomes, policy analysis and network analysis.

Student Outcomes

Determination of student outcomes is a possibility for determining effects on districts or states, but as mentioned previously, it is difficult to track the attribution of training-a-teacher-to-be-a-teacher-leader to state or district wide changes in student achievement. What would probably be most important in this type of analysis is to clearly explain the lack of a direct relationship between student outcomes and teacher leader training.

Policy Analysis

Perhaps more relevant to documenting the effects of teacher leader training programs is policy analysis. This sort of analysis can be conducted at either state or district levels. It is reasonable to expect that a teacher leader program and the leaders it produces would be in positions to affect policies. A policy analysis could also include an analysis of changes in state or local professional education organizations.

Network Analysis

The third possibility is network analysis. This procedure is designed to show the development and strength of the various networks of power and communication existing within a system. In this case, the system would be a district or state. Network analysis would allow the determination of the degree of teacher leaders' involvement or the involvement of institutions containing teacher leaders in the power and communication structures. It would also identify existing power structures and the relationships of these power brokers to the professional development effort.

Summary

In summary, evaluation is a complex undertaking that cannot be simply defined. There are many different interpretations of evaluation and no single correct approach to evaluation problems. Different approaches are designed to address different needs and different questions. Evaluators of teacher leader development programs need to carefully articulate their program's goals and objectives with reasonable, valued and documentable outcomes. Next, specific evaluation questions based on the interests and values of the stakeholders need to be developed. These questions will depend on the type of effects the program is expected to produce. For example, a program may be expected to deliver quality professional development. Therefore, evaluation questions would focus on the professional development sessions themselves. On the other hand, the program may be designed to produce statewide changes in the educational system. Then the evaluation questions would focus on changes in educational policy or delivery. Once determined, the evaluation questions should be matched with appropriate information gathering techniques. Then the data is collected and analyzed. The final step is providing the information in a manner that meets the needs

of the stakeholders, such as, state legislators, school superintendents, teachers, parents, and others. Providing useful data in appropriate formats is critical if the program is to survive and, if done well, can help the program to meet its goals.

References

Fetterman, D. (1989). *Ethnography step by step*. Newbury Park: Sage.

Fetterman, D. (1996). Empowerment evaluation: An introduction to theory and practice. In D. Fetterman, S. Kaftarian & A. Wandersman (Eds.), *Empowerment evaluation: Knowledge and tools for self-assessment and accountability*. Thousand Oaks: Sage.

Fetterman, D. (1997). Empowerment evaluation: A response to Patton and Scriven. *Evaluation Practice, 18 (3)*, 253-266.

Fraser, B. (1994). Research on classroom and school climate. In D.L. Gabel (Ed.), *Handbook of research on science teaching and learning*. New York: Macmillan.

Hestenes, D., Wells, M. & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher, 30*, 141-158.

Horizon Research, Inc. (1999). *Local systemic change 1998-1999 core evaluation collection manual*. Chapel Hill, North Carolina: Author.

House, E. (1983). Assumptions underlying evaluation models. In G. F. Madaus, M. Scriven & D. L. Stufflebeam (Eds.), *Evaluation models*. Boston: Kluwer-Nijhoff.

Howell, David C. (1987). *Statistical methods for psychology, second edition*. Boston: Duxbury Press.

Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards, second edition*. Thousand Oaks: Sage.

Kintsch, W., & van Dijk, T.A. (1978) Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.

Krueger, R. (1994). *Focus groups: A practical guide for applied research, second edition*. Thousand Oaks: Sage.

Lawrenz, F. (1986). Teacher to student transfer in energy education. *School Science and Mathematics, 86 (1)*, 1-9.

Louis, K., Marks H., & Kruse, S. (1996). Teachers' professional community in restructuring schools. *American Educational Research Journal, 33 (4)*, 757-789.

National Research Council. (1996). *National science education sandards*. Washington, DC: National Academy Press.

National Science Foundation. (1997). *User friendly handbook for mixed method evaluations*. Washington DC: National Science Foundation, Directorate for Education and Human Resources, Division of Research, Evaluation and Dissemination.

Newman F. & Associates. (1996). *Authentic achievement: Restructuring schools for intellectual quality*. San Francisco: Jossey-Bass.

Norusis, Marija J. (1994). *Advanced statistics 6.1*. Chicago: SPSS Inc.

Patton, M. (1997a). *Utilization-focused evaluation, New century edition*. Thousand Oaks: Sage.

Patton, M. (1997b). Toward distinguishing empowerment evaluation and placing it in a larger context. *Evaluation Practice, 18 (2),* 147-163.

Phi Delta Kappa National Study Commission on Evaluation. (1971). *Educational evaluation and decision making.* Itasca IL: F.E. Peacock.

Scriven, M. (1991). *Evaluation thesaurus, Fourth edition.* Newbury Park: Sage.

Stake, R. (1968). The countenance of educational evaluation. *Teachers College Record, 68,* 523-540.

Stevens, F., Lawrenz, F., & Sharp, L. (1993). *User friendly handbook for project evaluation: Science, mathematics, engineering and technology education.* Washington DC: National Science Foundation, Directorate for Education and Human Resources, Division of Research, Evaluation and Dissemination, GAO OCG-93-6TR.

Taylor, C., Fraser, B. & White, L. (1994, April). *CLES: An instrument for monitoring the development of constructivist learning environments.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans.

Trabasso, T., van den Broek, P., & Suh, S. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Processes, 12,* 1-25.

Worthen, B., Sanders, J. & Fitzpatrick J. (1997). *Program evaluation: Alternative approaches and practical guidelines, second edition.* White Plains, NY: Longman.

Yin, R. (1989). *Case study research design and methods.* Newbury Park: Sage.